Recognition-based Approach of Numeral Extraction in Handwritten Chemistry Documents using Contextual Knowledge

Nabil Ghanmi*†
*LORIA, Nancy, France
nabil.ghanmi@loria.fr
†eNovalys, Illkirch, France

Abdel Belaïd
Université de Lorraine - LORIA
Nancy, France
abdel.belaid@loria.fr

***Abstract*—This paper presents a complete procedure that uses contextual and syntactic information to identify and recognize amount fields in the table regions of chemistry documents. The proposed method is composed of two main modules. Firstly, a structural analysis based on connected component (CC) dimensions and positions identifies some special symbols and clusters other CCs into three groups: fragment of characters, isolated characters or connected characters. Then, a specific processing is performed on each group of CCs. The fragment of characters are merged with the nearest character or string using geometric relationship based rules. The characters are sent to a recognition module to identify the numeral components. For the connected characters, the final decision on the string nature (numeric or non-numeric) is made based on a global score computed on the full string using the height regularity property and the recognition probabilities of its segmented fragments. Finally, a simple syntactic verification at table row level is conducted in order to correct eventual errors. The experimental tests are carried out on real-world chemistry documents provided by our industrial partner eNovalys. The obtained results show the effectiveness of the proposed system in extracting amount fields.**

***Keywords*—numeral extraction; digit recognition; numeric string segmentation ; structural features; syntactic verification.**

I. Introduction

A chemistry document is a page taken from a laboratory notebook. It contains all the information needed to reproduce a performed chemical experiment. It is typically composed of a graphic illustration of the chemical reaction, a table containing the quantities of the used products and some text paragraphs describing the operating procedure. The segmentation of the document into regions was presented in our previous works [1] [2].

Since tables are worthy of attention for the information they contain, our work is currently focused there. More precisely, we investigate the extraction of the amount fields in these regions. This research is motivated by the fact that the amounts constitute the central pieces of the information contained in the tables. The location of such elements can be used to guide a table structure extraction work as already proposed in [3]. It can also help a subsequent recognition on the full region. The table is mostly composed of numeral values which can be mixed with text data (see Fig. I). The two symbols "=" (equal sign) and "," (decimal comma) may be present and their identification can be helpful.

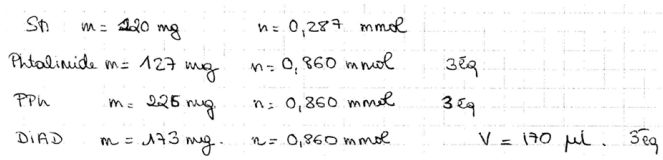


Fig. 1. Example of table taken from a chemistry document.

The extraction of the numerals is a challenging task since we are faced with several problems related to the handwritten nature of the writing such as the presence of fragmented characters and connected characters, the lack of a priori knowledge and the absence of any syntactic or physical constraints. In this paper we propose a recognition-based approach dealing with the above challenges. Being guided by contextual information, the fragments of the same character are merged and the connected digits are selected and segmented.

The rest of this paper is structured as follows: Section II provides an overall review of the existing methods for numeral extraction and segmentation. Section III presents an overview of the proposed system and explains each of its modules. Section IV reports the conducted experiments. Finally, some conclusions and future works are drawn in section V.

II. Related works

A. Numeral extraction

In [4], a method for extracting zip codes, phone numbers and customer codes from handwritten incoming mail documents is presented. The method does not use any segmentation or recognition. It works at line level by labeling each CC as digit, double-digit, separator or outlier using KNN classification based on some structural and contextual features. To find the best labeled sequence among the obtained sequences, a Markov based syntactical analyzer (modeling the syntax of the searched field) is applied on each line. This method works well but since no recognition is performed, some false alarms may occur in the digit class specially when the writing is hand-printed which is possible in our documents. This work is then extended [5] by adopting a segmentation-driven recognition strategy. Thus, each CC is considered successively as isolated, double or triple digit. At each time, a CC is segmented using the drop-fall algorithm and the obtained fragments are sent

to an MLP classifier for recognition. A 3-level recognition hypothesis is then obtained for each CC. The syntactic analysis is performed similarly as in [4] but considering the 3-level trellis instead of only one level. One of the most important steps in these methods is the syntactic analysis which allows a good verification and important corrections. This analysis is based on syntactic constraints on the extracted fields which is not the case of the amount fields in the chemistry documents.

An automatic numerical string extraction in handwritten, printed and mixed letters is proposed in [6]. Firstly, an explicit segmentation into characters is performed based on background skeletal graphs. The sequences of characters that are unlikely to be digits are excluded using geometric regularity measured on the heights, widths and interspacing of their constituent characters. Then, the decision whether a sequence is numeric or text is made using a likelihood of each character to be a digit (determined by neural network) and a likelihood of the entire sequence being a numeric or text string. This method presents a challenge concerning the segmentation into characters. We think that such segmentation still be a very hard task in handwritten document due to the high variability in size, shape and style (cursive or handprinted) of the script.

An older related work [7] is conducted for locating zip codes in handwritten address. The proposed method uses strong constraints on the spatial organization of the processed addresses as well as the form of the searched zip codes. The main idea of this work is to label each word in the address with a valid field (city, state, zip code, street number, etc.) based on its shape and the syntax constraints. This method does not suit our problem because we do not have enough a priori knowledge about the position neither the form of the amount fields.

B. Numerical string segmentation

The recognition of numeric strings is directly affected by the segmentation performance. In fact, this latter aims to single out the basic patterns which will be used in the recognition. A lot of efforts are devoted in this field. Sadri et al. [8] use a set of foreground and background features to construct possible segmentation paths. The foreground features represent the junction points in the outer contour of the investigated CC. The background features are the end points extracted on the skeleton of the vertical top and buttom profiles of the CC. A set of segmentation paths is generated by connecting the close feature points. Finally, a genetic algorithm is used to find the segmentation path with the highest segmentation/recognition confidence. In [9], a similar segmentation method is proposed but uses a verification step instead of the genetic algorithm. The verification is based on the combination of the outputs of the recognizer and two other classifiers used for checking over-segmentation and under-segmentation respectively. The segmentation method adopted in the two above works generates a large number of cutting paths which might be the cause of false alarms. For that matter, a robust verification or selection strategy is strongly needed.

In [10], a method based on the water reservoir is proposed. A water reservoir is a metaphor for a big top or buttom valleys. Firstly, the touching position is determined based on the position of the biggest water reservoir. Next, a set of rules based on the reservoir positions and sizes, close loop position and structural features are used to generate cutting path. This method works well for double digit segmentation but strings composed of more than two digits are not treated.

A "drop fall" segmentation algorithm is adopted in [5] to segment connected digits in handwritten mail documents. This method simulates the trajectory of an acid drop sliding vertically along the CC. When it reaches a valley, the drop crosses the CC contour. Among four possible paths (obtained by four possible movement directions: left/right combined with descending/ascending), the selection of the best one is performed based on the recognition confidence of the obtained fragments. These methods study only how to separate numerical strings which consist of two handwritten digits. In our system, we develop a segmentation method which deals with multi-numerical strings by estimating the number of the connected digits.

III. Proposed approach

A. System overview

The text within the tables in handwritten chemistry documents is mainly composed of isolated digits which have almost the same size. This information is used to estimate the digit dimensions. The flow-chart of the proposed system is illustrated in Fig. 2. After being extracted, the CCs are grouped into 3 groups based on their dimensions. The small CCs represent character fragments and thus must be merged with other CCs in order to constitute complete character(s). The CCs of dominant size are isolated characters. They will be sent to a digit recognizer in order to decide if they are digits or not. The big CCs represent strings of either letters or digits. Among these CCs, those which are evidently non-digits based on their shapes are excluded. The others will be segmented in order to recognize their fragments separately and a final decision on their nature (numerical or non-numerical strings) is made.

B. Connected component analysis

This module aims to extract the CCs in the table area, identify the two special symbols "," and "=" and cluster the other components into 3 groups according to their dimensions. Then, each group will be sent to a specific treatment. Let $\{C_i\}$ be the set of the CCs and $\{bb_i\} = \{(x_i, y_i, w_i, h_i)\}$ their bounding boxes. Let y_{bl} be the y-coordinate of the base line of a table row. The decimal comma is identified as a CC satisfying the following conditions: 1) its biggest part is situated below the base line, 2) is vertically elongated and 3) does not exhibit any fluctuation. These conditions are expressed by the following equations

$$\begin{cases} y_{cc} + \alpha.w_{cc} \leq y_{bl} & (1) \\ h_{cc}/w_{cc} > r_{min} & (2) \\ \max\limits_{x}(cross\_count(CC, x)) < 2, \ x \in [\![x_{cc}, x_{cc} + w_{cc}]\!] & (3) \end{cases}$$

where $cross\_count(CC, x)$ is the number of times a vertical ray of x-coordinate x crosses the CC, α and r_{min} are two parameters experimentally fixed to 0.2 and 1.5 respectively.
Similarly, the equal symbol is identified as a CC pair

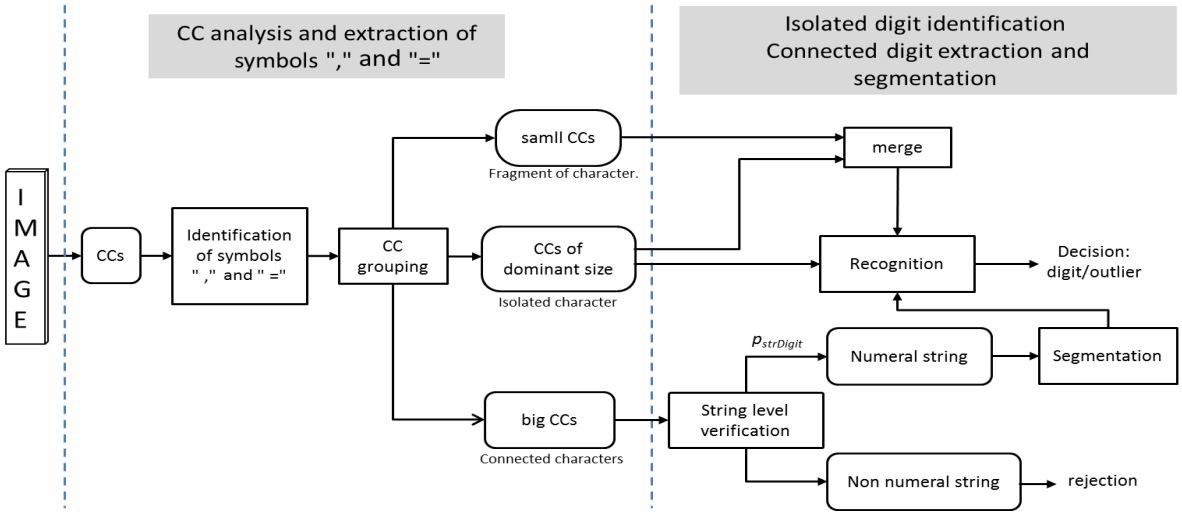


Fig. 2. Flow-shart of the scheme.

(CC_1, CC_2) satisfying the conditions expressed by the following equations:

$$\begin{cases} O_h(CC_1, CC_2) > 0 & (4) \\ h_{cc_i}/w_{cc_i} > r_{max},\ i = 1, 2 & (5) \end{cases}$$

where $O_h(CC_1, CC_2)$ represents the horizontal overlapping value between C_1 and C_2.

For the CC clustering, the width and the height histograms of the CCs are created. Let h_d and w_d be respectively the height and width that correspond to the peak in the associated histogram. The set of CCs are clustered into three groups as illustrated in Fig. 3.

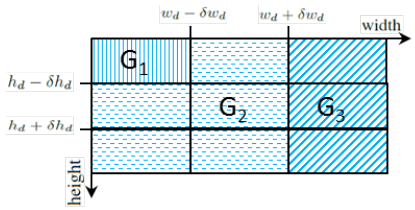


Fig. 3. 3 groups of CC according to their dimensions. G_1: small CCs, G_2: dominant CCs and G_3: big CCs.

Generally, a small CC corresponds to a character fragment. It must be merged with another component to form an integer number of characters. This operation is performed based on the horizontal overlapping and the vertical distance between the CCs. The CCs of group G_2 correspond to isolated characters (digits or letters) and those of G_3 are words, fragment of words or connected digits. Among all the CCs, those which have a descender part (w.r.t. the base line) might be non-numerics and then are discarded. The processing of the CCs of group G_2 and G_3 are respectively explained in section III-C and III-D.

C. Isolated digit identification

According to [12], MLPs and SVMs normally give better performance in handwritten recognition than other classifiers with the same features. Therefore, due to the efficiency in implementation, we use a support vector machine (SVM) for the recognition of isolated digits.

Firstly, the CC images are size normalized to fit in a 20×20 pixel box while conserving their aspect ratio. Then, these images are translated to position the mass center of the pixels at the center of 28×28 box. For the recognition, the binary image itself is given as input to the SVM.

All CCs of group G_2 are potential digits but non-digits are also present. To deal with this problem, we adopt a two-stage strategy: a one-class classification designed for digit/outlier discrimination and a classification into 10 classes used for the discrimination between the digits. Here, the use of one-class classification is justified by the fact that the outliers present a large variability. Thus, it is difficult to collect representative data of outliers to be used in the training if we want to use a classification into two classes . Digits, that constitute the data of the target class, are collected from the studied documents. A one-class SVM (kernel = RBF, $\gamma = 0.1$, $\nu = 0.05$) is used for this classification. A set of 4500 digits is used to train this classifier.

After this first stage, the components identified as digits are submitted to the second SVM to be recognized. In this stage, a rejection is also possible. A CC is rejected if the maximum a posteriori probability is below a threshold [11]. Once the isolated digits are identified, a mean width w_m is estimated on all those digits. This value is used to help locating the zones of connections in numerical stings.

D. Numerical string identification

In this section, we assume that a CC of group G_3 can be an alphabetic or a numerical string. Thus we do not deal with mixed strings where letters and digits are connected. This assumption is based on the observation on the handled documents where connections between numbers and letters are rarely encountered.

1) Preliminary filtering: A CC composed of touching digits exhibits a height regularity throughout its width. A preliminary filtering based on this postulate aims to exclude some evident

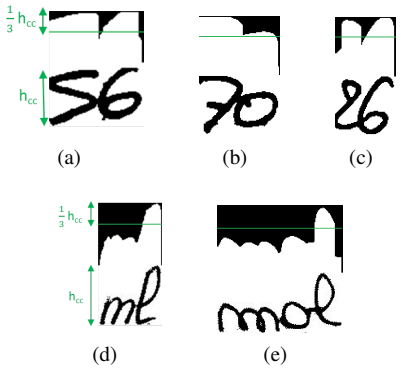


Fig. 4. Upper profile of numerical strings (a, b and c) and non-numerical strings (d and e).

non-numerical strings. For this purpose, we define $score_{str}$ as a score based on the height regularity of the full string (without any segmentation) that expresses to what degree the CC of the string looks like a numerical string.

To evaluate the height regularity, the upper profile is used. It shows the height variation, as illustrated in Fig. 4.

Let lv be the sum of the length of the valleys which are below $h_{cc}/3$ and let lv_m be the ratio lv/w_{cc}. The score $score_{str}$ computed on a given CC is defined in function of its lv_m. This function must satisfy the following axiom: the more regular the height of the CC is (i.e. lv_m is low), the higher score to be a numerical string is. Particularly, if the height is regular everywhere along the width of the CC ($lv_m = 0$), this latter keeps a complete chance to be numeral and thus it receives a value 1 as a score. If the half (or more) of the CC has a height below $h_{cc}/3$, then the CC has a score 0. To better characterize the scoring function, we determine statistically its value when $lv_m = 2/3$. For this value of lv_m, $score_{str}$ is 0.5. We choose to model this function by a polynomial. It is given in equation 6.

$$score_{str}(lv_m) = \begin{cases} 0 & \text{si } lv_m \geq 0.5 \\ -3lv_m^2 - \frac{1}{2}lv_m + 1 & \text{si } lv_m \leq 0.5 \end{cases} \quad (6)$$

The function curve of the equation 6 is plotted in Fig. 5.

The string score is computed for each CC. The CCs that have a score 0 are considered as non numerals and then are discarded. The other CCs are sent to a segmentation module for additional verification.

2) String segmentation and recognition: Using the estimated width w_m of the isolated digit (identified in section III-C), the touching position is situated in the zone horizontally

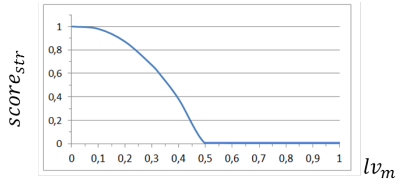


Fig. 5. Curve of the score function.

centered on the points of x-coordiante $x_{cc}+i*w_m$, where $i = 1, ..., n$ and $i * w_m < w_{cc}$. For illustration, see Fig. 6. Unlike [10] where the CC is firmly divided to guide the touching position search, which is only applicable on two connected digits, we are based on contextual information (the estimated w_m) to localize the touching area and thus to estimate the number of digits.

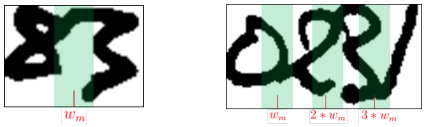


Fig. 6. Touching zone (in green) estimated based on isolated digit width.

The adopted segmentation method is based on the one proposed in [8]. But instead of considering all junction points in the outer contour of the CC and all end points of the skeletonized upper and buttom profile, only those situated on the touching zone are considered. This reduces significantly the number of characteristic points and then facilitates the decision making for the construction of the cutting path. For illustration, see Fig. 7.

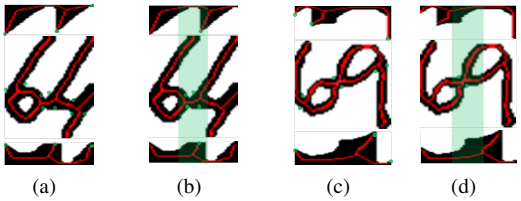


Fig. 7. Feature points extracted on the outer contour of the CC and on the skeleton of the upper and buttom profile. The skeleton is drawn in red. (a) and (c) illustrate all feature points of the numeral strings 64 and 69 respectively. (b) and (d) illustrate only the feature points situated in the estimated connection zone of these two strings.

Once the segmentation is performed, the obtained fragments $\{f_j\}_{j=1..n}$ are fed to the digit recognizer in the same way as the isolated digits. Let $p_{indiv}(f_j\ is\ num)$ (here, the index "$indiv$" stands for "individual") be the probability of the first proposition of the recognizer applied to the fragment f_j seen individually. The final probability $p(f_j\ is\ num)$ of f_j being numeric is determined as the individual probability p_{indiv} weighted by the score computed on the string, i.e., $p(f_j\ is\ num) = p_{indiv}(f_j\ is\ num) \times score_{str}$. If this probability is greater than a threshold (set experimentally to 0.5 here), the fragment is considered as numeric. Finally, the full string is considered as numeric if the majority of its fragments are numeric. The string value is given by the concatenation of its fragment values.

E. Postprocessing

As our symbol (“,” and “=”) identification is considerably accurate (see Table III), we are based on the position of the identified symbols to make a simple verification and eventually correct some potential errors in numeral extraction. These two simple rules are used for this verification:

1) The CC situated just at the left of the symbol ”=” is non-numeric and the one situated at the right is numeric

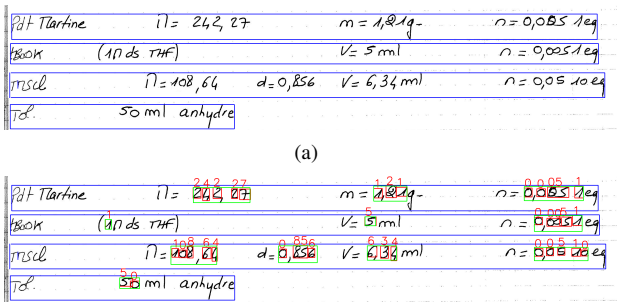
2) The two CCs situated at the left and the right of the symbol "," are numeric

To perform the verification, a left to right scan of the CCs in each table row is performed and corrections based on the above rules are made.

Once identified, the numeral CCs in the same line are finally grouped based on inter-bounding box distances to constitute amount fields. The grouping is based on classical techniques and can be briefly summarized as follows: two adjacent numeral CCs are grouped together if the distance between their bounding boxes is below a threshold (estimated using the horizontal distance histogram) [2]. Also, two numeral CCs separated by a decimal comma are grouped together.

IV. Experimental results

Experiments are conducted on a dataset of 215 documents provided by our industrial partner eNovalys[1] . Each document contains a table, making a total of 4294 non-empty cells. 2606 are data cells and thus contain amount fields (the others are header cells). For each document, the input of the system is the table region defined by the set of its rows. The output is the bounding boxes and the values of the numeral fields contained in these rows (Fig. 8).



(a)

(b)

Fig. 8. Example of (a) the input and (b) the output of the system. The extracted numeral fields are bounded in green. The numeral CCs and their values are in red.

A. Global evaluation

A coarse evaluation regarding the final goal, namely numeral field extraction, is performed. For this purpose, the tables are annotated at word level using GEDI. A word is described by its bounding box and its content. This annotation allows the discrimination between numeral fields and other data (chemical measure or unit, product name, etc.) based on the content of the word. The obtained results are reported in Table I.

TABLE I. Amount field extraction accuracy

# amount fields	# detected	# Corrected	Recall(%)	Precision(%)
2606	2729	2118	81, 27	77, 61

The correctness of a detected amount field in comparison with the ground truth is determined based on the intersection-over-union measure used in [13]. An amount field is considered correctly detected if:

$$O(A_G, A_D) = \frac{2|A_G \cap A_D|}{|A_G| + |A_D|} \geq 0.9 \qquad (7)$$

[1] www.enovalys.com

where $|A_G|$ and $|A_D|$ denote the area of the ground-truth and the detected amount respectively and $|A_G \cap A_D|$ denotes the area of the intersection of their bounding boxes.

For a best investigation of the error origins and subsequently the limits of the proposed method, a finer evaluation of each of the main steps of the proposed system is performed. For this purpose, an additional semi-automatic annotation of all CCs within the tables is performed. Each CC is described by its bounding box and its content. Some statistics on the CCs are given in Table II

TABLE II. Statistics on the CCs within the tables

CC		total number
Special symbol	decimal comma	1517
	equal sign	128
Isolated Character	numeral	7925
	non-numeral	6769
Numerical string	double	661
	triple	81
	quadriple	2
Non-numerical string		2281

B. CC extraction and special symbol identification

The extraction rates of the special symbols (Fig. 9) are summarized in Table III.

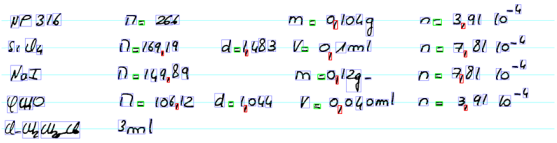


Fig. 9. Illustration of CC extraction and special symbol identification. The comma symbol is squared in red, the equal sign in green and other CCs in blue.

TABLE III. Special symbol extraction rates

Symbol	Precision (%)	Recall (%)
=	91,76	95,12
,	90,21	95,18

We note that the precision of decimal comma extraction is slightly lower than other rates. This is mainly due to the detection of some subscripts (frequently present in the name of chemical products) as comma symbols.

C. Isolated digit identification

A set of 4500 digits, equitably distributed over the 10 digit classes ($\sim$ 450 samples/class), is used for the training of both one-class SVM and 10-class SVM. Then, each CC of group G_2 is sent to these trained classifiers to be recognized. The main objective here was the identification of isolated digits and not determining the exact values of the digits. Thus, the performance of this module is measured in terms of correctly identified digits. Table IV shows the obtained performance on all the isolated characters.

TABLE IV. Isolated digit identification rates

# CCs	# digits	# detected	# FP (FP_{rate})	# FN (FN_{rate})
14694	7925	7796	239(3, 53%)	368(4, 64%)

The false positive rate FP_{rate} and the false negative rate FN_{rate} are given by the following formulas:

$$FP_{rate} = \frac{FP}{FP+TN} \; ; \quad FN_{rate} = \frac{FN}{TP+FN}$$

The main errors occur when the writing is handprinted. They are due to the confusion between some letters and digits such as “O” and “0”, “l” and “1”, “S” and “5”, etc.

D. Numercial string identification

The main objective of the filtering step is to exclude as many non-numerical strings as possible while retaining all numerical strings. The evaluation of this step is performed considering this aspect. Our filtering step achieves a rejection rate of $42,5\%$ of non-numerical strings while accepting $99,01\%$ of numerical strings. It is true that the false acceptance rate is high, but this is not a serious problem since this kind of error can be recovered later during the segmentation/recognition step. Some false acceptance cases are illustrated in Fig. 10. On the other hand, a very low percentage of numerical strings

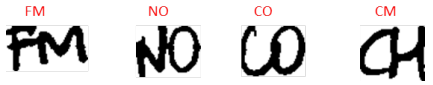


Fig. 10. Some false acceptance cases.

are rejected. This is an important advantage of this filtering since a false rejection is not recoverable. Some false rejection cases are presented in Fig. 11. The numerical strings (in the top of the image) do not exhibit a regular height along their widths and thus they are rejected by the filtering step. We think that such confusion are unavoidable especially when we see that they are very similar to non-numerical strings (in the buttom of the image).

The proposed segmentation method is evaluated separately on all the numerical strings. For each document, the estimated mean width w_m of the isolated digits and the numerical strings are sent to the segmentation module. The obtained results are summarized in Table V

TABLE V. NUMERICAL STRING SEGMENTATION ACCURACY

numerical strings	Total number	Correctly segmented
double	661	$95,16\%$
triple	81	$81,48\%$ $17,28\%$ partially correct
quadriple	2	1 correct 1 partially correct

A visual analysis reveals that most unsuccessful segmentations are due to a big overlap between the digits. Other errors are also noticed in the strings containing the digit “1” in the form of stick.

V. CONCLUSION

A complete procedure for amount field extraction in tables of handwritten chemistry documents is presented in this paper. The proposed system adopts a recognition-based strategy where different treatments such as CC merging, filtering, segmentation and recognition are performed. These treatments use contextual information extracted on the CCs within the table. This system is evaluated on real-world documents taken

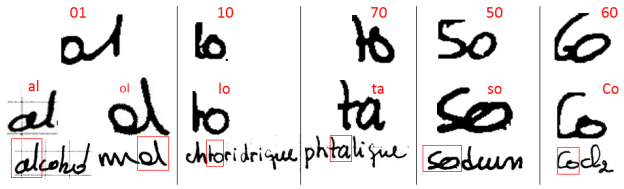


Fig. 11. Some false rejection cases. The numerical strings (in the top) are rejected

from chemistry laboratory notebook. The obtained results are promising. Although the system has been implemented for extracting amount fields in handwritten documents of chemistry, we believe that it can be applied to extract numeral fields in other kinds of documents (such as statistical reports) provided that some a priori information (especially related to the size and the frequency of the digits) are available. Concerning our future work, we are currently working on preparing a consistent ground truth in order to enlarge the dataset used for the experimentation. We also plan to integrate the numeral extraction results in our previous work [3] focusing on table structure extraction.

REFERENCES

[1] N. Ghanmi and A. Belaïd, *Extraction de formules chimiques dans des documents manuscrits composites*, Colloque International Francophone sur l’Ecrit et le Document, pp. 185-197, 2014.

[2] N. Ghanmi and A. Belaïd, *Table Detection in Handwritten Chemistry Documents Using Conditional Random Fields* , International Conference on Frontiers in Handwriting Recognition, pp. 146-151, 2014.

[3] N. Ghanmi and A. Belaïd, *Table Structure Extraction in Handwritten Chemistry Documents*, International Conference on Document Analysis and Recognition, pp. 296-300, 2015.

[4] G. Koch, L. Heutte and T. Paquet, *Automatic extraction of numerical sequences in handwritten incoming mail documents*, Pattern Recognition Letters, vol. 26, no. 8, pp. 1118-1127, 2005.

[5] C. Chatelain, L. Heutte and T. Paquet, *Segmentation-Driven Recognition Applied to Numerical Field Extraction from Handwritten Incoming Mail Documents*, International Workshop on Document Analysis System, pp. 564-575, 2006.

[6] M. M. Haji, T. D. Bui, C. Y. Suen, *Automatic extraction of numeric strings in unconstrained handwritten document images*, Document Recognition and Retrieval, pp. 2-10, 2011.

[7] S.N. Srihari and E.J. Keubert, *Integration of handwritten address interpretation technology into the united states postal service remote computer reader system*, International Conference on Document Analysis and Recognition, pp. 892-896, 1997.

[8] J. Sadri, C. Y. Suen, T. D. Bui, *A genetic framework using contextual knowledge for segmentation and recognition of handwritten numeral strings*, Pattern Recognition, vol. 40, no. 3, pp. 898-919, 2007.

[9] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, *Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy*, Pattern Analysis and Machine Intelligence, vol. 24, no. 11, pp. 1438-1454, 2002.

[10] U. Pal, A. Belaïd and Ch. Choisy, *Touching numeral segmentation using water reservoir concept*, Pattern Recognition Letters, vol. 24, pp. 261-272, 2002.

[11] C.L.Liu, H. Sako, H. Fujisawa, *Performance evaluation of pattern classifiers for handwritten character recognition*, International Journal on Document Analysis and Recognition, vol. 4, no. 3, pp. 191-204, 2002.

[12] C.L. Liu, K. Nakashima, H. Sako and H. Fujisawa, *Handwritten digit recognition: benchmarking of state-of-the-art techniques*, Pattern Recognition, vol. 36 n. 10, pp. 2271-2285, 2003.

[13] F. Shafait and R. Smith, *Table Detection in Heterogeneous Documents*, International Workshop on Document Analysis System, pp. 65-72, 2010.